Text Binarization in Color Documents

Efthimios Badekas, Nikos Nikolaou, Nikos Papamarkos

Department of Electrical and Computer Engineering, Image Processing and Multimedia Laboratory, Democritus University of Thrace, 67100 Xanthi, Greece

Received 18 July 2006; revised 14 February 2007; accepted 14 March 2007

ABSTRACT: This article presents a new method for the binarization of color document images. Initially, the colors of the document image are reduced to a small number using a new color reduction technique. Specifically, this technique estimates the dominant colors and then assigns the original image colors to them in order that the background and text components to become uniform. Each dominant color defines a color plane in which the connected components (*CC*s) are extracted. Next, in each color plane a *CC* filtering procedure is applied which is followed by a grouping procedure. At the end of this stage, blocks of *CC*s are constructed which are next redefined by obtaining the direction of connection (DOC) property for each *CC*. Using the DOC property, the blocks of *CC*s are classified as text or nontext. The identified text blocks are binarized properly using suitable binarization techniques, considering the rest of the pixels as background. The final result is a binary image which contains always black characters in white background independently of the original colors of each text block. The proposed document binarization approach can also be used for binarization of noisy color (or gray-scale) document images. Several experiments that confirm the effectiveness of the proposed technique are presented. © 2007 Wiley Periodicals, Inc. Int J Imaging Syst Technol, 16, 262–274, 2006; Published online in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/ima.20092

Key words: color quantization; text localization; binarization; segmentation; document processing

I. INTRODUCTION

Interest about exploiting text information in images and video has grown notably during the past years. The ability of text to provide powerful description of the image content, the convenience of distinguishing it from other image features and the provision of extremely important information, reasonably attracts the research interest. Content-based image retrieval, OCR, page segmentation, license plate location, address block location, and compression, are some examples based on text information extraction from various types of images.

A main categorization of text identification methods include texture based techniques (Jain and Bhattacharjee, 1992; Jain and Zhong, 1996) and connected components (*CC*s) based techniques (Fletcher and Kasturi, 1988; O'Gorman, 1993; Chen and Chen, 1998; Sobottka et al., 2000; Hase et al., 2001; Strouthopoulos et al., 2002). Some hybrid approaches have also been reported in the literature (Zhong et al., 1995; Jung and Han, 2004). Texture based techniques are time consuming and use character size restrictions. The main advantage is their capability of detecting text in low resolution images. On the other hand, *CC*s based techniques are fast and exploit the fact that characters are segmented.

Most approaches for text identification refer to gray or binary document images. Only recently, some techniques have been proposed for text identification and extraction in color documents. Strouthopoulos et al. (2002) proposed a method for text extraction in complex color documents. It is based on a combination of an adaptive color reduction technique and a page layout analysis approach, which uses a Kohonen SOM neural network in order to identify text blocks. Zhong et al. (1995) presented a hybrid system for text localization in complex color images. According to this system, a color segmentation stage is performed by identifying local maxima in the color histogram. Heuristic filters on the *CC*s of the same color plane are applied and noncharacter components are removed. A second approach based on local spatial variance, which locates text lines, is also proposed. Chen and Chen (1998) proposed a method for text block localization on color technical journals cover images. Initially, the colors of the image are reduced using a YIQ color model based algorithm. With the Sobel operator and through a binarization process, strong edges are isolated. Primary blocks are then detected with the Run Length Smearing Algorithm and finally classified with the use of nine features that underlie on fuzzy rules. Sobottka et al. (2000) proposed an approach to extract text from color documents and journal covers. The image is quantized with an unsupervised clustering method and the text regions are then identified combining a top-down and a bottom-up technique. An algorithm for character string extraction from color documents is presented by Hase et al. (2001). First the number of representative colors of a document is determined. Potential character strings are then extracted from each color plane using multistage relaxation. When all extracted elements are superimposed, a strategy which utilizes the likelihood of a character string and a conflict resolution is followed to produce the final result. A detailed review on text information extraction techniques is presented by Jung et al. (2004).

Correspondence to: Prof. Nikos Papamarkos; e-mail: papamark@ee.duth.gr

Contract grant sponsor: TEI Serron

Most of the complete systems for text binarization include, as a first stage, the identification of the text regions. This stage is necessary to apply the binarization procedure locally (in each text block), where the contrast between the characters and the background colors is usually very high. Strouthopoulos and Papamarkos (2000) proposed a system for multithresholding of gray-scale documents. Recently, Wang et al. (2005) proposed a color text image binarization technique based on a color quantization procedure and a binary texture analysis. However, this technique neglects the identification of the text region stage that is does not include a page layout analysis technique. Thillou and Gosselin (2005) proposed a color binarization technique for complex camera-based images based on wavelet denoising and a color clustering with K-means. However, in this approach the technique does not include any text extraction stage and is applied only for document images containing already detected text.

The proposed technique is based on a document color reduction procedure and a text localization technique proposed by Nikolaou and Papamarkos (2006) and Nikolaou et al. (2006), respectively. The color reduction procedure is firstly applied and reduces the colors to a small number in accordance to the content and the dominant colors of the document image. This technique is based on an edge preserving smoothing algorithm and a mean-shift procedure (Comaniciu and Meer, 2002). After the application of this powerful color reduction procedure, we have uniform background and characters with solid colors. The suitable reduction of the document image colors is crucial for the effectiveness of the entire localization technique. The goal of this paper is to propose a new technique for the binarization of the identified text blocks of color document images. The proposed technique overcomes the difficulties associated with mixed type color documents such as complex color cover pages. Specifically, for this type of color documents, text and graphics are highly mixed with the background and even more, in many cases, the background cannot be defined. To handle varying colors of the text in the image, a combination strategy is implemented among the binary images (we call them color planes) obtained by the color reduction procedure. Specifically, after color reduction, each color defines a color plane in which the connected components (*CC*s) are extracted and a *CC*s filtering and grouping procedure is applied. At the end of this stage, blocks of *CC*s are constructed, which are next refined by obtaining the direction of connection (DOC) property for each *CC*. Next, the blocks of *CC*s are classified as text or nontext blocks using their DOC property. The text blocks identified in the different color planes are superimposed and the text identification of the entire document is achieved. To have uniform colors in the presentation of the extracted text blocks, their color is further reduced to only two colors using the adaptive color reduction (ACR) technique proposed by Pamarkos (1999) and Papamarkos et al. (2002). Finally, the color text blocks are converted to black and white, that is, black characters in white background independently of the original colors of each text block. This final binarization stage is very beneficial for the cases where the characters of the identified text blocks must be recognized through an OCR application. That is, the identified text blocks, especially the color ones, cannot be fed into OCR tools for recognition unless appropriate binarization has been carried out.

The proposed technique consists of the following main stages:

Stage 1. Color reduction (Section II).

Stage 2. Connected component filtering (Section III).

Stage 3. Initial elements grouping (Section IVA).

Stage 4. Final elements grouping and blocks formation (Section IVB).

Stage 5. Classification of blocks and color planes superimposition (Section V).

Stage 6. Binarization of the identified text blocks (Section VI).

The proposed technique performs satisfactory in the majority of mixed type color documents even in the cases where the processing documents have horizontal and/or vertical text orientation with about 15° of angle tolerance. The proposed technique is implemented in visual environment and it has been extensively tested with success with a large number of color documents.

II. COLOR REDUCTION

In the first stage of the proposed technique, the colors of the processing document image are suitably reduced to a small number in accordance to the content and the dominant colors of the document image. Based on these colors, the initial image is decomposed into color planes each one of which contains only the pixels that have a dominant color. This color reduction technique was proposed by Nikolaou and Papamarkos (2006). The purpose is to create a simplified version of the color document image where characters can be extracted as solid items, by utilizing a connected component analysis and a labeling procedure.

The overall process consists of the following stages:

Stage 1. Edge preserving smoothing.

Stage 2. Color edge detection.

Stage 3. RGB color space approximation (subsampling).

Stage 4. Initial color reduction.

Stage 5. Mean shift.

Generally, a color reduction algorithm for text information extraction applications based on connected components, must be able to perform its task without oversegmenting characters and still preventing fusion with the background. Additionally, it is desirable to merge low contrast objects with their background and create large compact areas. This will result to a small number of connected components, so the outcome of a text information extraction algorithm will be extensively improved. The color reduction technique that is adopted in this article satisfies these criteria. A brief description of the color reduction technique used is given in this section.

A. Edge Preserving Smoothing. To enhance the quality of the resulting image (especially for noisy document images) and to improve the performance of the method, an edge preserving smoothing filter is used as a preprocessing step. First, in a 3 × 3 spatial window the Manhattan color distances from the center pixel are calculated:

$$d_i = \left|R_{a_c} - R_{i_i}\right| + \left|G_{a_c} - G_{i_i}\right| + \left|B_{a_c} - B_{i_i}\right| \qquad (1)$$

The coefficients for the convolution mask of the filter are extracted as follows:

$$c_i = (1 - d_i)^p \qquad (2)$$

That is, c_i receives larger values for smaller values of d_i. This concludes to the following convolution mask:

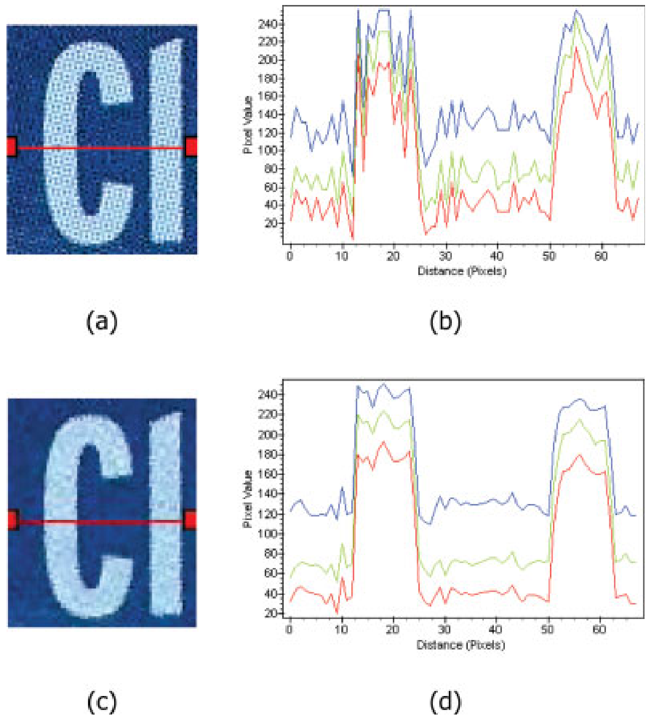


Figure 1. The effect of the edge preserving smoothing filter on a color document. (a) Original noisy color document, (b) RGB pixel profile of line $y = 44$ on the original document. (c,d) Filtered document ($p = 10$) and the pixel profile of the same line. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\frac{1}{\sum_{i=1}^{8} c_i}\begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & 0 & c_5 \\ c_6 & c_7 & c_8 \end{bmatrix} \qquad (3)$$

Factor p scales exponentially the color differences. Thus, it controls the amount of blurring performed on the image. As it gets larger, coefficients with small color distance from the center pixel increase their relative value difference from coefficients with large color distance, so the blurring effect decreases. A fixed value 10 is used for all experiments. The center pixel of the convolution mask is set to zero to remove impulsive noise. Figure 1 shows the effect of the filter on a color document. It can be noticed that noise is reduced without affecting edge points.

B. Color Edge Detection. Next, based on a simple methodology, color edge detection is performed. With the use of the Sobel operator, the edge strength for each one of the three color channels is calculated. The final edge value is obtained by choosing:

$$G(x,y) = \max\left\{|G^r(x,y)|, |G^g(x,y)|, |G^b(x,y)|\right\} \qquad (4)$$

where $|G^r(x,y)|$, $|G^g(x,y)|$, $|G^b(x,y)|$ the edge values for red, green, and blue channel, respectively.

C. RGB Color Space Approximation (Subsampling). From the edge image $G(x,y)$, a representative set of samples is chosen and the 3D histogram of the RGB color space is constructed. Specifically, the image is subsampled by selecting only those pixels that are local minima in the eight-neighborhood on the edge image. The resulted set of pixels has the following useful properties:

Edge points are not represented in this set so fuzzy areas are avoided.

Spatially, the samples are always inside the objects of the image.

Every object's color is represented in the sample set.

Figure 2 shows an example of approximating the color distribution according to this subsampling technique. As it can be observed, the selected pixels are placed very close to the cluster centers of the initial image's RGB distribution. Thus, it is assumed that every member of the extracted set of samples can be considered as a candidate cluster center. This assumption is used in the next stage to initially reduce the colors.

D. Initial Color Reduction. The color reduction procedure continues by choosing a random RGB sample s_i from the previous obtained set of samples and estimates the initial color classes according to the following adaptive procedure:

Step 1. Define a cube with length of side $2h_1$. Considering $s_i = (r_i,g_i,b_i)$ as the center of the cube, calculate a new point $s_{m_i} = (r_{m_i}, g_{m_i}, b_{m_i})$ where r_{m_i}, g_{m_i}, b_{m_i} the mean values of red, green, blue channels, respectively in the defined cube.

Step 2. Label all points contained in the cube that has been examined.

Step 3. Choose randomly a new unlabeled sample and go to *Step 1*. If all the samples are labeled the procedure stops.

The new set of points S_m is used to initially reduce the colors of the image. This is done by assigning to the pixels of the original image the color of their nearest neighbor (Euclidean distance) in S_m. The size of S_m (number of points) depends on the size of the cube, namely on h_1. With the use of value 32, the number of the obtained colors is relatively large and usually smaller than 100, thus the resulted image is oversegmented.

E. Mean Shift. The color centers obtained from the initial color reduction step are used by a mean shift operation (Comaniciu and Meer, 2002) to locate the final points of the RGB color space. Based on the final color centers, the algorithm extracts the final result.

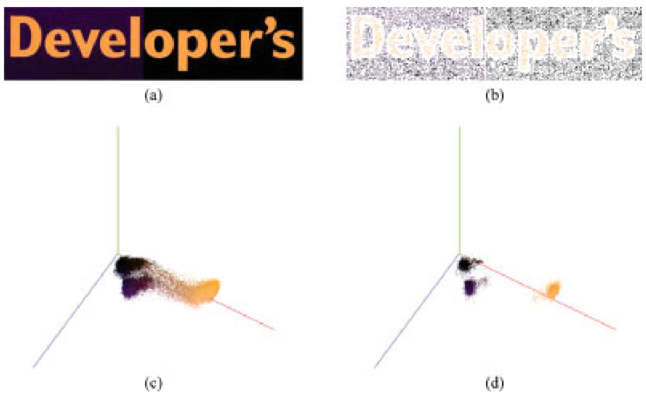


Figure 2. (a) Original color document, (b) local minima pixels, (c) RGB color distribution of (a), (d) RGB color distribution of local minima pixels. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

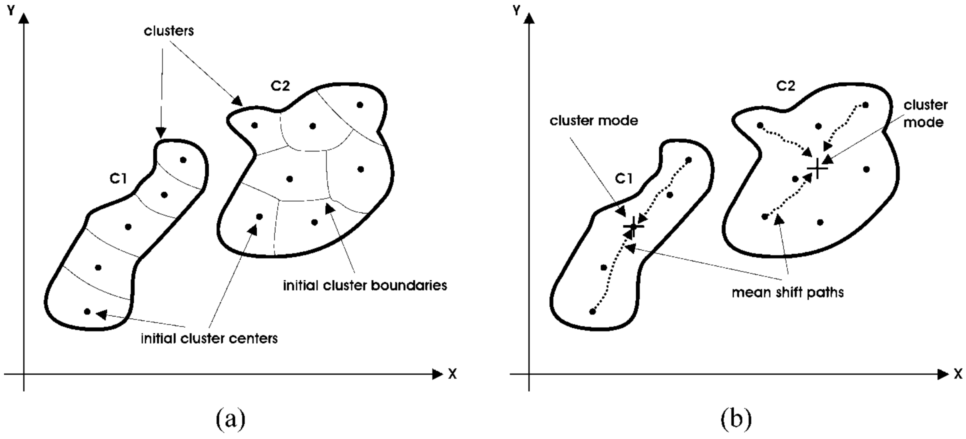


Figure 3. Hypothetical case of clustering in the 2D space, (a) the two randomly shaped clusters C1 and C2 are initially oversegmented, (b) the final result is adopted by mean shifting the initial cluster centers (mode detection).

Adopting this approach, it is possible to deal with a clustering problem where the clusters are randomly shaped. The initial color reduction and the mean shift operations are graphically depicted in Figure 3.

III. CONNECTED COMPONENTS PROPERTIES AND FILTERING

In each color plane, connected components (*CC*s) are identified. The enclosing rectangle of a connected component (*CC*) is defined as its bounding box. Let CC_i be a connected component. Every *CC* is characterized by the following set of features:

i. $BB(CC) = \{W_i, H_i\}$ is the bounding box of CC_i. W_i represents the width and H_i the height of the CC_i.
ii. $psize(CC_i)$. The number of pixels of the CC_i.
iii. $bsize(CC_i)$. The size of $BB(CC_i)$.
iv. $dens(CC_i) = psize(CC_i)/bsize(CC_i)$. The density (or saturation) of CC_i.
v. $elong(CC_i) = \min\{W_i, H_i\}/\max\{W_i, H_i\}$. The elongation of CC_i.

According to the above features, CC_i is considered as a nontext object if

i. $psize(CC_i) < T_{\text{psize}}$, where T_{psize} is taken equal to 6 pixels.
ii. $dens(CC_i) < T_{\text{dens}}$, where T_{dens} was set at 0.08. This means that CC_i must cover no less than the 8% of the $BB(CC_i)$ in order to consider it as text object.
iii. $elong(CC_i) < T_{\text{elong}}$, where T_{elong} was set at 0.08. This means that the width W_i of a character element cannot be 12.5 times larger than H_i (and the opposite).
iv. $BB(CC_i)$ contains more than T_{e1} *CCs* having the same color with CC_i. This check is used to classify as nontext, large objects (background) that contain in their bounding boxes smaller objects with the same color. It is found that $T_{\text{e1}} = 3$, is a proper value as there is not any character consisted of more than three objects. For example the Greek character "Θ" consists of two objects.

It is noticed that the thresholds have been carefully selected after several tests in order not to reject character elements. This filtering procedure can be considered as a preprocessing step and is focused only on removing very noisy or large elements resulted from the color reduction procedure. In addition, it speeds up the entire text identification technique since the number of *CC*s significantly decreases.

IV. BLOCK FORMATION

A. Initial Grouping of Connected Components. Let CC_i be a connected component. To connect it with another *CC*, a region must be defined around it and only the objects whose centroids are included in this region are possible to be connected with CC_i. Figure 4 depicts the specific region $R(CC_i)$, which is defined as the set of all pixels (x_i, y_i) satisfying the following inequality:

$$d_{\min} \leq d_i \leq d_{\max} \tag{5}$$

where d_i represents the Euclidean distance of pixel (x_i, y_i) from the centroid of CC_i, $d_{\max} = c_{\text{d}} \max\{W_i, H_i\}$,and $d_{\min}$ a small constant value (usually 5 pixels). Coefficient c_{d} adjusts the maximum size of $R(CC_i)$. In our experiment c_{d} was taken equal to 3 resulting to neighborhood regions that contain all necessary information that specify if a *CC* belongs to an horizontal or vertical structure block.

Only objects whose centroid is located inside $R(CC_i)$ are considered as objects that it is possible to be connected with CC_i. For a connection to be established between a pair of *CC*s (CC_i, CC_j), the following similarity criterion, based on *psize*, must also stand:

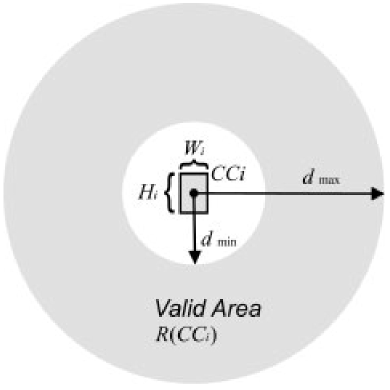


Figure 4. Definition of the neighborhood region $R(CC_i)$ of CC_i.

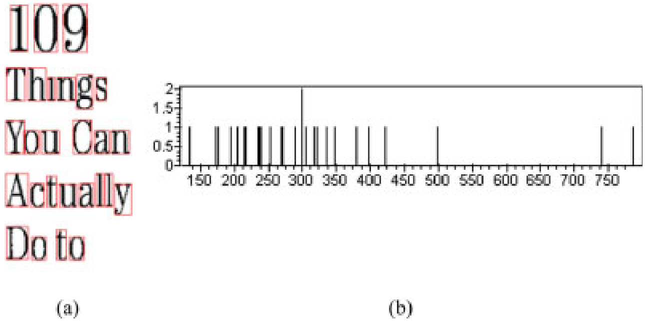


Figure 5. (a) Clean text image with significant character variations, (b) size histogram of (a). The smaller character has *psize* $=$ 134 and the largest *psize* $=$ 786 (5.87 size ratio). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$$\frac{\max\{\mathrm{psize}(\mathrm{CC_i}),\mathrm{psize}(\mathrm{CC_i})\}}{\min\{\mathrm{psize}(\mathrm{CC_i}),\mathrm{psize}(\mathrm{CC_j})\}} \le T_{\mathrm{psr}} \tag{6}$$

T_{psr} takes the large value 7 in order *CC*s having *psize* ratio larger than this threshold value not to be connected. Tests that were contacted on clean text images of various fonts showed that character elements have size ratio differences no more than 6.5 even in the case of joined characters, so the value of 7 was chosen to have an additional safety tolerance. Figure 5 shows such an example where the *psize* ratio was found to be equal to 5.87.

The final constraint for *CC*s connection is related to a distance measure defined in Simon et al. (1997). This distance measure is adopted but it is used in a different way. The horizontal block distance (HBD(*i,j*)) and the vertical block distance ((VBD(*i,j*)) for the pair of CC_i and CC_j is defined as:

$$\mathrm{HBD}(i,j) = \max\{Xl_i, Xl_j\} - \min\{Xr_i, Xr_j\} \tag{7}$$

$$\mathrm{VBD}(i,j) = \max\{Yl_i, Yl_j\} - \min\{Yr_i, Yr_j\} \tag{8}$$

Equations (7) and (8) are graphically depicted in Figure 6. Note that if HBD(*i,j*) < 0, CC_i and CC_j overlap in the vertical direction and if VBD(*i,j*) < 0 (as in Fig. 6) they overlap in the horizontal direction. Having this in mind, we state that if

$$\mathrm{HBD}(i,j) \ge 0 \ \wedge \ \mathrm{VBD}(i,j) \ge 0 \tag{9}$$

a connection between CC_i and CC_j is not possible. In words, Eq. (9) indicates that no overlapping occurs in any of the two directions. It is noticed that connections are bidirectional, which means that connection conditions must apply both for CC_i towards CC_j and the opposite. This can be seen clearly in Figure 7, where character "y" connects with character "n" but not "n" with "y," and this leads the two characters not to be connected.

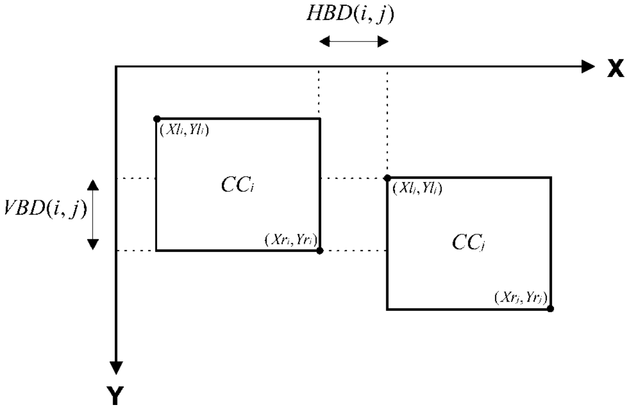


Figure 6. Definition of *HBD*(*i,j*) and *VBD*(*i,j*) between two *CC*s.

The result of the initial grouping process is the creation of connection sets associated to each *CC*.

$$C(CC_i) = \{c_1^i, \ldots, c_{cn}^i\} \tag{10}$$

where *cn* the number of connections for each CC_i. If $C(CC_i) = \emptyset$ then no match exists in $R(CC_i)$. Any *CC* having this property is considered as a nontext object (isolated). Thereby, further filtering of nontext objects is achieved. An example of the initial grouping procedure is shown in Figure 8c. In most cases, characters are assigned with more than four connections. This helps in gathering more information about *CC*s than taking into account only the closest neighbors.

B. Assigning the Direction of Connection Property to Connected Components. Based on the results of initial grouping, we proceed to the characterization of *CC*s with a property named direction of connection (DOC). The purpose of this strategy is to create homogenous blocks of elements, that is, to spatially discriminate text from nontext elements in order for the classification stage (Section V) to be able to classify text blocks from nontext blocks. To define the DOC property, the following two metrics are introduced:

$$H_o(CC_i) = -\sum_{j=1}^{cn} \mathrm{VBD}(i,j) \tag{11}$$

$$V_o(CC_i) = -\sum_{j=1}^{cn} \mathrm{HBD}(i,j) \tag{12}$$

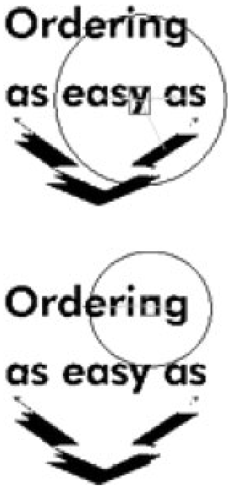


Figure 7. Experimental example where connections between *CC*s are shown. The region $R(CC_i)$ of character "y" contain the centroid of character "n" but the two characters are not connected because the centroid of character "y" is out of the region $R(CC_j)$ of character "n."

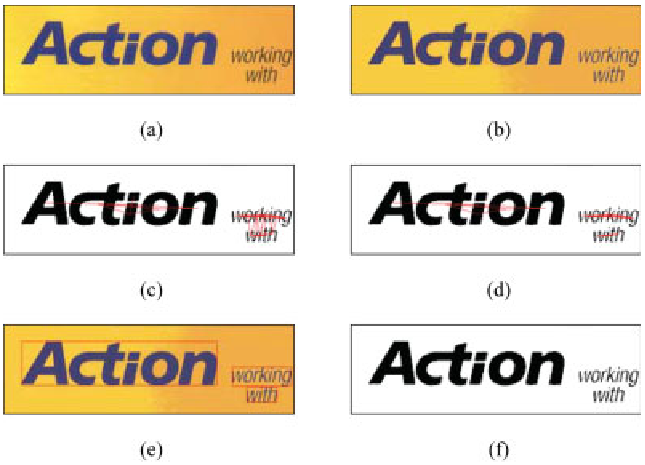


Figure 8. (a) Original color document, (b) color reduction result (four colors), (c) initial grouping, (d) false connections removal, (e) blocks identified after superimposition of all color planes, (f) binarization of text block. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

$H_o(CC_i)$ and $V_o(CC_i)$ measure the total amount of CCs overlapping in horizontal and vertical direction, respectively. The DOC property is defined as follows:

$$DOC(CC_i) = \begin{cases} 1, & \text{if } (H_o > V_o) \wedge \left(\dfrac{H_o}{V_o}\right) > T_o \wedge (H_o \geq H_i) \\ 2, & \text{if } (H_o < V_o) \wedge \left(\dfrac{V_o}{H_o}\right) > T_o \wedge (V_o \geq W_i) \end{cases} \tag{13}$$

$DOC(CC_i)$ =1 indicates that CC_i belongs to a horizontal structure and $DOC(CC_i) = 2$ that CC_i belongs to a vertical structure. T_O is set at 2 and represents the overlapping difference between horizontal and vertical direction that a CC must have in order to be characterized. It should be noticed that for a tolerance of about 15° (in either horizontal or vertical direction) a significant amount of overlapping between CCs remains and thus these blocks can also be characterized.

The application of the earlier procedure does not categorizes all the CCs to an horizontal or vertical structure. That is, there are still CCs with DOC value equal to zero. To characterize more CCs, an additional step is applied. According to this step, for each CC that have DOC value equal to zero, the DOC values of the two nearest CCs, which are connected with the processing CC, are compared. If these two DOC values are equal then the processing CC gets the same DOC value.

Finally, before the classification of the blocks, all false connections between CCs are removed. A false connection is defined as follows:

1. Connection between two CCs that still have DOC values equal to zero.
2. Connection between an element with DOC value equal to 1 or 2 (valid element) and an element with DOC value equal to zero (zero element). If the zero element does not belong to the two closest elements connected with the valid element then the connection between these two elements is considered as false connection.
3. If a CC_i belongs to a horizontal or vertical structure ($DOC(CC_i) = 1$ or $DOC(CC_i) = 2$) then every connection with a CC_j that has no vertical or horizontal overlapping ($VBD(i,j) \geq 0$ or $HBD(i,j) \geq 0$) respectively, is considered as false connection.

After the application of the earlier procedure text blocks are spatially discriminated from nontext blocks and the final block classification stage can be applied. Figure 8d shows the result obtained after the false connection removal. As it can be seen, text elements are grouped in the sense of text lines.

V. CLASSIFICATION OF BLOCKS

In this final stage, a classification procedure is applied which classifies the text and nontext blocks. Because of the fact that text block elements are very likely to be assigned with DOC values 1 or 2, a statistical metric is used to reflect this. Let B_j be a structure block containing N CCs, and

$$BH_j = \{CC_i \in B_j : DOC(CC_i) = 1, \quad i = 1, \ldots, k\} \tag{14}$$

$$BV_j = \{CC_i \in B_j : DOC(CC_i) = 2, \quad i = 1, \ldots, m\} \tag{15}$$

represent the subset containing all the CC_i of the structure block B_j that have $DOC(CC_i) = 1$ and $DOC(CC_i) = 2$, respectively. The entire structure block B_j is considered to be a horizontal text block if

$$\frac{k}{N} \geq T_{\mathrm{p}} \tag{16}$$

and a vertical text block if

$$\frac{m}{N} \geq T_{\mathrm{p}} \tag{17}$$

where $T_{\mathrm{p}} \in [0.5, 0.9]$. From our experiments, we have found that a suitable value for T_{p} is 0.75. As the value of T_{p} increases, the possibility to have correct text block characterization is increased. However, a high value of T_{p} may lead to some missed text blocks. The blocks that contain CCs whose DOC value is different (1 and 2) are named mixed blocks. From these blocks the BV_j subset of CCs is removed if $k > m$ and the $k < m$ subset is removed if BH_j. If $k = m$ then these blocks are not considered as text blocks.

The blocks obtained in the different color planes of the image are superimposed and all the blocks that are included in larger blocks are removed. This can be done because their regions will be binarized in the next stage, during the binarization process of the larger block. It is obvious that by removing these unnecessary blocks the binarization process (Section VI) is performed faster. The identified text blocks of Figure 8a are shown in Figure 8e.

VI. BINARIZATION OF TEXT BLOCKS

The final binary image is obtained by the application of the ACR technique locally to the regions of all identified text blocks. It is noticed that in most of these regions, the contrast between the characters and background colors is very high. The goal of the binarization procedure is to achieve the best possible segmentation between characters and local background. In most cases, the color

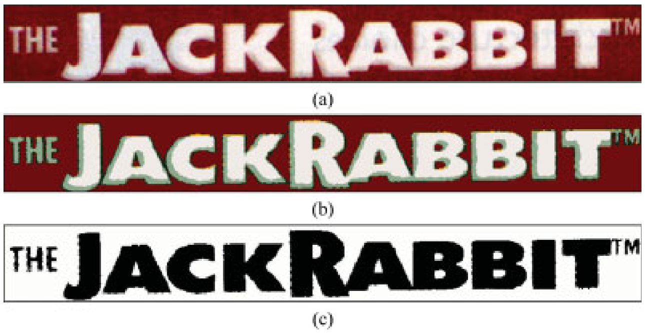


Figure 9. (a) Original color document, (b) color reduction result, (c) binarization result. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

reduction technique, because of the shadow appearing in the characters edges produces locally noisy pixels (Fig. 9b). The presence of these pixels does not induce the proposed technique to identify the text blocks. However, to overcome this effect and to achieve proper presentation of the detected characters a local binarization procedure is applied to each identified text block. Figure 9c depicts the result obtained after the application of the local binarization procedure.

The entire image region, which does belong to any identified text block, is considered as background and gets white color. The text blocks regions are converted to a binary form by applying locally the ACR technique. Specifically, according to the ACR technique, a Kohonen SOM neural network, consisting of an input and an output layer, is locally applied to each text block. The input layer has three neurons and the output layer only two. The Kohonen SOM is fed by the initial RGB color components and leads, after training, to only two dominant colors. Next, these two colors are converted to black and white according to their Euclidian distance form them (Fig. 9c). To achieve uniformity in the final document image, all text blocks are converted to have black characters in white background. To do this, in each binary text block a text block color inversion procedure described in the next section is applied.

A. Text Block Color Inversion. A document image can contain text blocks constituted of characters whose colors are lighter than the colors of local background. In these cases, the binary blocks obtained by the binarization procedure will have white characters in black background and the colors of these blocks must be inverted (Figs. 11f and 13f). That is, in most of the cases it is preferable to have uniform background. For this reason, in the binary text regions obtained from the previous stages, a background estimation procedure is applied. According to this procedure, the colors of the text blocks having white characters in a black background are inverted to have finally, in the entire document, black characters in a white background. The decision for the inversion is taken for each binary text block as follows:

Step 1. The white and black Run Length Histograms are calculated in the horizontal and vertical directions of the binarized block region. Consider as $R_{\mathrm{B}}(i)$, $i = 1, \ldots, M_{\mathrm{B}}$ the "Black" Run Length Histogram where M_{B} is the longest run of black pixels and R_{W} (i), $i = 1, \ldots, M_{\mathrm{W}}$ the "White" Run Length Histogram where M_{W} is the longest run of white pixels.

Step 2. The maximum longest run of these two histograms corresponds to the background color. If $M_{\mathrm{B}} \leq M_{\mathrm{W}}$, then the specific text block has already black characters in white background, and there is no need for color inversion. If $M_{\mathrm{B}} > M_{\mathrm{W}}$ then the colors of the specific text block are inverted.

B. Document Binarization Techniques. Binarization techniques, which have been proposed especially for document binarization, can alternatively be used for the binarization of the identified text blocks. These techniques are preferable to be applied in cases where the text blocks contain noise that must be removed. A recent description of the five document binarization techniques included in our system is given as follows:

1. Bernsen's technique (1986) uses a local threshold which is calculated as follows:

$$T(x,y) = \begin{cases} \dfrac{P_{\mathrm{low}} + P_{\mathrm{high}}}{2}, & \text{if } P_{\mathrm{high}} - P_{\mathrm{low}} \geq L \\ GT, \text{ if } P_{\mathrm{high}} - P_{\mathrm{low}} < L \end{cases} \tag{18}$$

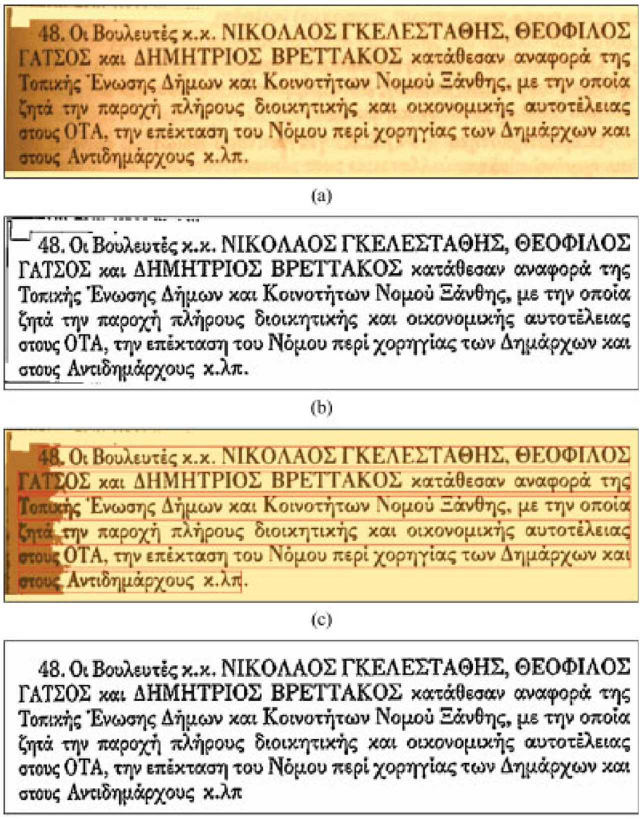
   where P_{low} and P_{high} are the lowest and the highest gray-level value in a $N \times N$ window centered in the pixel (x, y), respectively and GT a global threshold value (for example a threshold value that is calculated from the application of the method of Otsu to the entire image).
2. Niblack's technique (1986) which also uses a local threshold calculated as follows:

$$T(x,y) = m(x,y) + ks(x,y) \tag{19}$$

   where $m(x,y)$ and $s(x,y)$ are the local mean and standard deviation values in a $N \times N$ window centered on the pixel (x,y), respectively.
3. Sauvola and Pietikainen's technique (2000) proposes the following calculation for the local threshold:

$$T(x,y) = m(x,y)\left[1 + k\left(1 - \frac{s(x,y)}{R}\right)\right] \tag{20}$$

   where $m(x, y)$ and $s(x, y)$ are the same as in the previous technique and R is a constant equal to 128 in most cases.
4. The adaptive logical level technique (Yang and Yan, 2000) (ALLT) is based in the initial calculation of the stroke width (SW) of the characters of the processing document image. Then for each pixel calculates a local threshold taking into account information of pixels placed in a $SW \times SW$ window, centered on the processing pixel. An analytical description of this technique and specific improvements are proposed by Badekas and Papamarkos (2003). It is noticed that the initial detection of text blocks and the independently application of the ALLT in each text block, leads to local estimation of the *SW* of the characters for each text block. This, in contrast to the work of Yang and Yan (2000) where a global *SW* is used, leads to superior binarization results.
5. The improvement of integrated function algorithm (Trier and Taxt, 1995) (IIFA) is based on the initial calculation of the edges of the characters which are represented of a pair of pixels labeled "− +" or "+ −". Next, the regions surrounded by edges are filled and finally, false characters are removed. An analytical description of this technique is given by Badekas and Papamarkos (2003).



48. Οι Βουλευτές κ.κ. ΝΙΚΟΛΑΟΣ ΓΚΕΛΕΣΤΑΘΗΣ, ΘΕΟΦΙΛΟΣ
ΓΑΤΣΟΣ και ΔΗΜΗΤΡΙΟΣ ΒΡΕΤΤΑΚΟΣ κατάθεσαν αναφορά της
Τοπικής Ένωσης Δήμων και Κοινοτήτων Νομού Ξάνθης, με την οποία
ζητά την παροχή πλήρους διοικητικής και οικονομικής αυτοτέλειας
στους ΟΤΑ, την επέκταση του Νόμου περί χορηγίας των Δημάρχων και
στους Αντιδημάρχους κ.λπ.

(a)

48. Οι Βουλευτές κ.κ. ΝΙΚΟΛΑΟΣ ΓΚΕΛΕΣΤΑΘΗΣ, ΘΕΟΦΙΛΟΣ
ΓΑΤΣΟΣ και ΔΗΜΗΤΡΙΟΣ ΒΡΕΤΤΑΚΟΣ κατάθεσαν αναφορά της
Τοπικής Ένωσης Δήμων και Κοινοτήτων Νομού Ξάνθης, με την οποία
ζητά την παροχή πλήρους διοικητικής και οικονομικής αυτοτέλειας
στους ΟΤΑ, την επέκταση του Νόμου περί χορηγίας των Δημάρχων και
στους Αντιδημάρχους κ.λπ.

(b)

48. Οι Βουλευτές κ.κ. ΝΙΚΟΛΑΟΣ ΓΚΕΛΕΣΤΑΘΗΣ, ΘΕΟΦΙΛΟΣ
ΓΑΤΣΟΣ και ΔΗΜΗΤΡΙΟΣ ΒΡΕΤΤΑΚΟΣ κατάθεσαν αναφορά της
Τοπικής Ένωσης Δήμων και Κοινοτήτων Νομού Ξάνθης, με την οποία
ζητά την παροχή πλήρους διοικητικής και οικονομικής αυτοτέλειας
στους ΟΤΑ, την επέκταση του Νόμου περί χορηγίας των Δημάρχων και
στους Αντιδημάρχους κ.λπ.

(c)

48. Οι Βουλευτές κ.κ. ΝΙΚΟΛΑΟΣ ΓΚΕΛΕΣΤΑΘΗΣ, ΘΕΟΦΙΛΟΣ
ΓΑΤΣΟΣ και ΔΗΜΗΤΡΙΟΣ ΒΡΕΤΤΑΚΟΣ κατάθεσαν αναφορά της
Τοπικής Ένωσης Δήμων και Κοινοτήτων Νομού Ξάνθης, με την οποία
ζητά την παροχή πλήρους διοικητικής και οικονομικής αυτοτέλειας
στους ΟΤΑ, την επέκταση του Νόμου περί χορηγίας των Δημάρχων και
στους Αντιδημάρχους κ.λπ

(d)

Figure 10. (a) Original image, (b) binarization result obtained by the application of the IIFA to the entire image, (c) the reduced color image (three colors) with the identified text blocks (red blocks), (d) binarization result obtained by the application of the IIFA to the grayscale versions of the identified text blocks. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

The Otsu's technique (1979), which is based on the calculation of a global threshold, can also be used to decrease the computational cost. Also, to improve further the binarization results, the final step of the IIFA can be applied as a post processing step.

It is noticed that the above techniques must be applied to the gray-scale versions of the identified text blocks. To binarize properly text blocks whose colors must be inverted, there is necessary to apply the above binarization techniques to the negative gray-scale version of these blocks. For this reason, the decision for the color inversion in each text block is taken before the application of a document binarization technique. The gray-scale version of each block is binarized using the Otsu's technique and the obtained binary image is checked for inversion as described in Section VIA. If there is need for inversion then the proper binarization technique is applied to the negative gray-scale version of the specific text block.

In contrast to the global case, the local application of the above binarization techniques leads to superior binarization results. This can be easily observed in the example shown in Figure 10. The specific document image is coming from the old Greek Parliamentary Proceedings. The binary image shown in Figure 10d does not contain background noise in contrast to the binary image shown in Figure 10b. The first image is obtained by applying the IIFA only to the pixels of the identified text blocks (Fig. 10c), considering the rest of image pixels as background. The second image is produced by the global application of the IIFA technique with the same parameters and as it is observed, this approach leads to the presence of noisy pixels. Experiment 3, gives another example for the application of Bernsen's technique to the identified text blocks of a color book cover image.

VII. EXPERIMENTAL RESULTS

To test the proposed technique, a large image database of color documents was created (1000 images at 150–300 dpi). Some images were scanned from color book covers and journals and some were obtained from the Internet. The proposed technique was extensively tested with different types of color documents.

To measure the performance of our algorithm, the text block precision rate (BPR) was used.

$$\mathrm{BPR} = \frac{N_c}{N_e} \qquad (21)$$

where N_c are the correct extracted text blocks and N_e the total number of extracted blocks. We have found a mean value of BPR $\approx 85\%$.

A. Experiment 1. In this experiment, the color document image shown in Figure 11a is used. Firstly, the color reduction technique described in Section II is applied to the original image producing an image with five colors that is shown in Figure 11b. Figure 11c depicts the result of the initial grouping stage on the red color plane that contains both text and nontext *CC*s. Note that some formed blocks contain text and nontext elements. Also, for some *CC*s no match was found so these are removed from further examination. The final grouping stage, as described in Section IVB, is shown in Figure 11d. The blocks of the initial grouping procedure are refined to form the final blocks that will be classified with the use of the information obtained by the DOC of each *CC*. The final extracted text blocks, from all color planes, considered as text are shown in Figure 11e. The binarization of the identified text blocks obtained by the application of the ACR technique is shown in Figure 11f. In this image, there are text blocks that need color inversion. Thus, the final binarization result obtained, after the inversion stage, is shown in Figure 11g. The size of the document image is 829 × 1171 pixels and the processing time in a Pentium 4 CPU 2.4 GHz is 11.23 s.

To evaluate the ability of the proposed technique to estimate text regions, it is compared with the results obtained by a commercial OCR application. Figure 11h presents the text blocks identified by applying Fine Reader 7. In contrast to the results obtained by the proposed technique, Figure 11h has many missed text blocks.

The binarization results obtained by the proposed technique are also compared with the results obtained by other binarization techniques. Figures 12a and 12b present the results of the techniques of Otsu and Sauvola which represent the global and local binarization category, respectively. The proposed technique is also compared with the IIFA and ALLT that take into account not only the image gray-scale values, but also the structural characteristics of the characters. The results obtained by these techniques are shown in Figures 12c and 12d respectively. In contrast to the result obtained by the proposed technique (Fig. 11g), the results in Figure 12 have many lost characters and others that are not binarized properly.

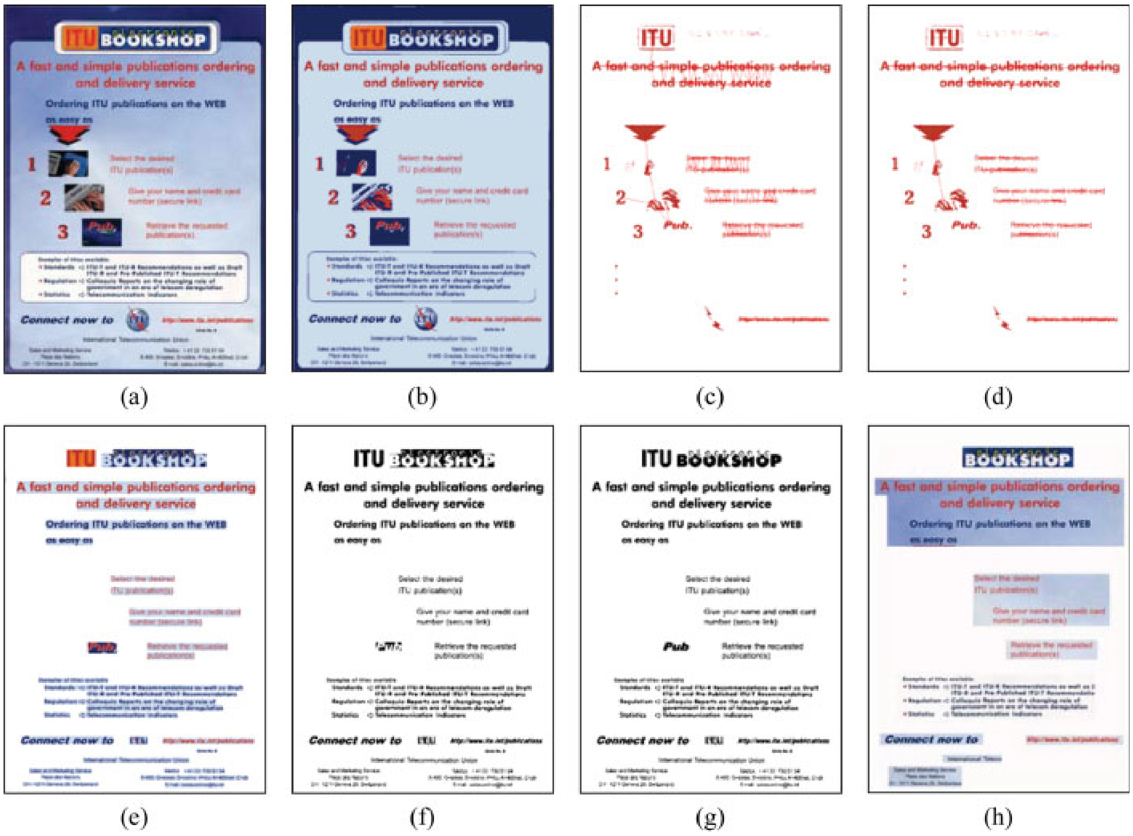


Figure 11. (a) original image, (b) original image after color reduction (five colors), (c) initial elements grouping for red color plane, (d) final elements grouping, (e) all identified text blocks after superimposition of all color planes ($T_p = 0.75$), (f) binarization of the identified text blocks without colors inversion, (g) binarization result with all characters converted to have black color, (h) text blocks obtained by using a commercial OCR application. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

B. Experiment 2. The second experiment presents the application of the proposed technique to the color book cover document shown in Figure 13a. It is noticed that this document image contains both vertically and horizontally aligned text and also, the skewed word "BONUS." The white color plane that is selected to demonstrate the stages of the proposed technique contains all of the above alignments of the text. The initial grouping of the *CC*s is shown in Figure 13c while Figure 13d shows the connections between the *CC*s after the false connection removal stage. The identified text blocks are shown in Figure 13e. Figures 13f and 13g present the binarization results before and after the color inversion stage, respectively. It can be seen that the proposed technique successfully extracts all the text blocks independently of their orientations. In the specific experiment the binarization technique of Otsu, which

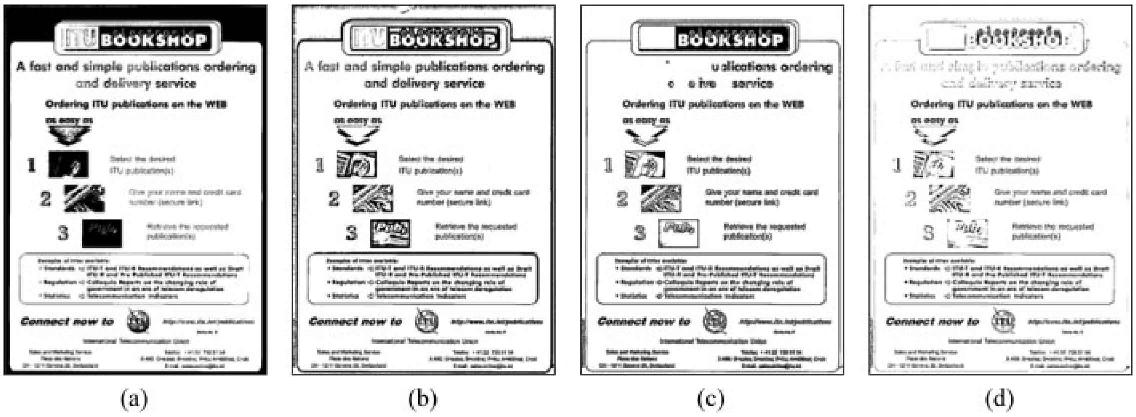


Figure 12. The results obtained by applying to the document of Figure 11(a) the binarization techniques of (a) Otsu, (b) Sauvola, (c) IIFA and (d) ALLT.

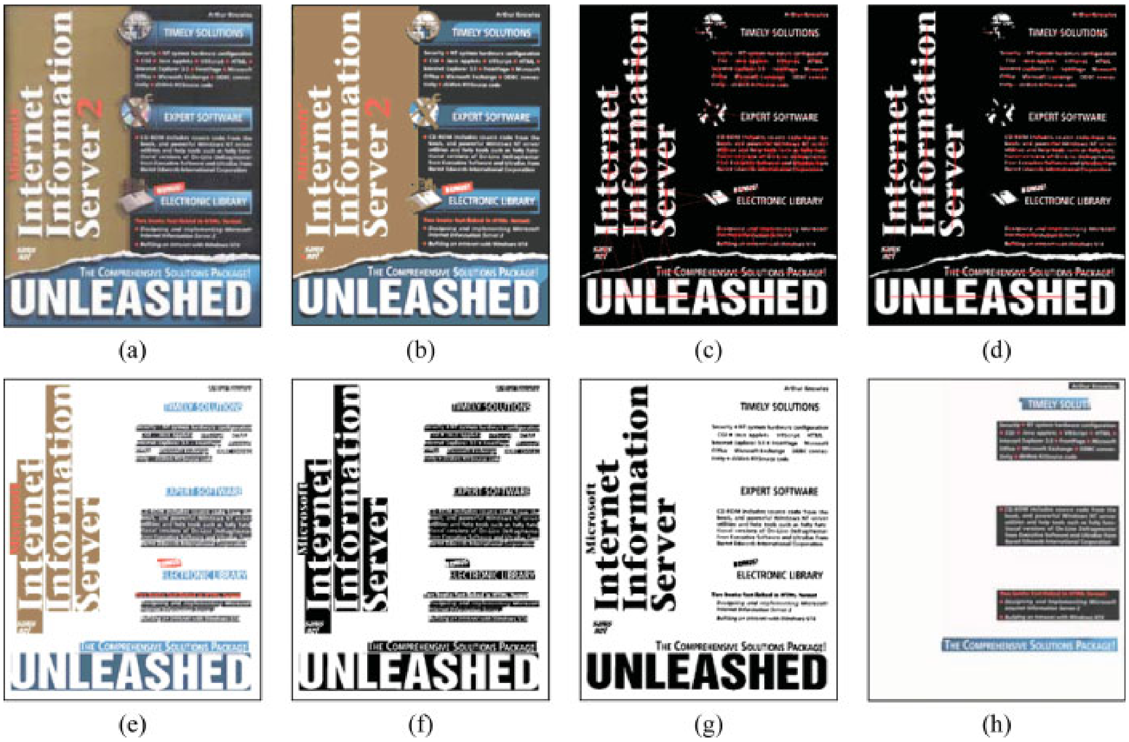


Figure 13. (a) original image, (b) original image after color reduction (seven colors), (c) initial elements grouping for white color plane, (d) final elements grouping, (e) all identified text blocks after superimposition of all color planes ($T_p = 0.75$), (f) binarization of the identified text blocks without colors inversion, (g) binarization result with all characters converted to have black color, (h) text blocks obtained by using a commercial OCR application. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

has the minimum computational cost, is used to binarize the identified text blocks in the final stage. The processing time in a Pentium 4 CPU 2.4 GHz is 0.82 s, while the size of the document image is 648 × 813 pixels.

Figure 13h shows the text blocks obtained by the commercial application where all vertical and many horizontal text blocks are not identified. Figure 14 presents the results obtained by the four other binarization techniques. It can be easily seen that all of these techniques failed to conclude to a binary image with black characters in white background in contrast to the result obtained by the proposed technique (Fig. 13g).

C. Experiment 3. This experiment analyzes the use of a document binarization procedure in the final stage of the proposed technique and the benefits obtained. Figures 15b–15d show the results obtained in each stage, as in the previous experiments. In Figure 15e, it can be seen that there are some small blocks that wrongly considered as text. There are also large blocks having many colors. The binarization of these blocks, by a global binarization technique, leads to the presence of noisy pixels. Figure 15f shows the result obtained using the ACR technique for the binarization of all identified text blocks (the inversion stage is included in this result). It can be seen that there are some noisy pixels especially

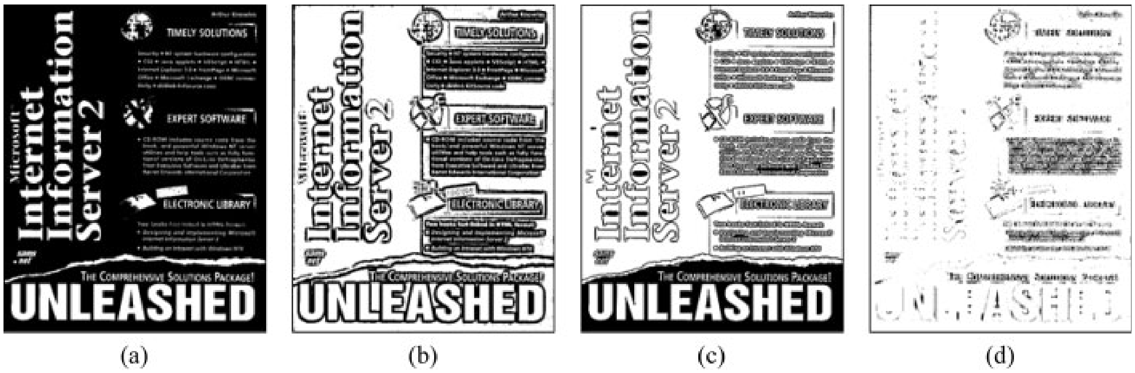


Figure 14. The results obtained by applying to the document of Figure 13(a) the binarization techniques of (a) Otsu, (b) Sauvola, (c) IIFA and (d) ALLT.

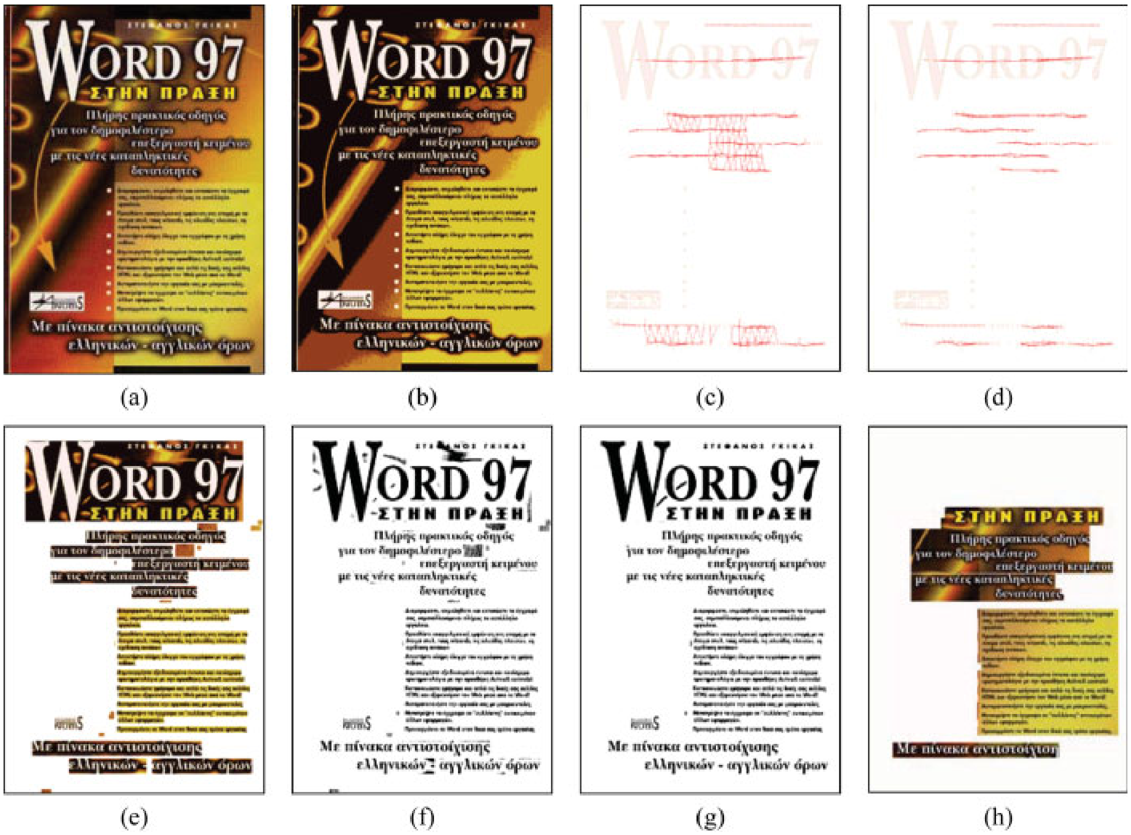


Figure 15. (a) original image, (b) original image after color reduction (seven colors), (c) initial elements grouping for white color plane, (d) final elements grouping, (e) all identified text blocks after superimposition of all color planes ($T_p = 0.75$), (f) binarization of the identified text blocks using the ACR technique, (g) binarization of the identified text blocks using the Bernsen's technique ($N = 7$ and $L = 40$) (h) text blocks obtained by using a commercial OCR application. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

in the large text block placed at the top of the document image. The quality of the obtained image is also reduced by the wrongly identified text blocks which are binarized and included in the final result. Figure 15g depicts the binarization result obtained by the application of Bernsen's technique to each identified text block (the inversion stage also included). The technique is applied with size of the neighboring window $N = 7$ and the parameter L equal to 40. As it can be seen all the wrongly identified text blocks are removed as false printed objects and also, the binarization of the corrected identified text blocks are much better than the ACR binarization. The size of the document image is 971 × 1376 pixels and the processing time in a Pentium 4 CPU 2.4 GHz is 8.37 s. Figure 15h shows the text blocks obtained by the commercial application. Figure 16 present the results obtained by the four other binarization techniques.

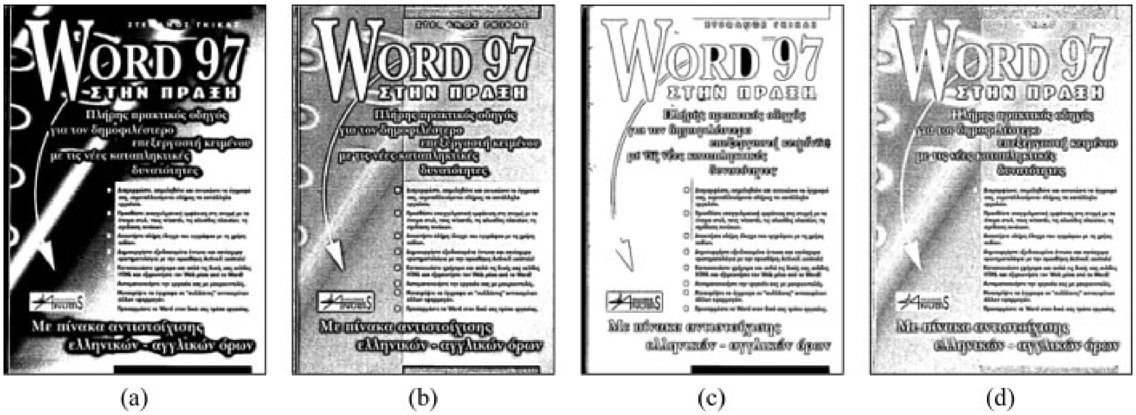
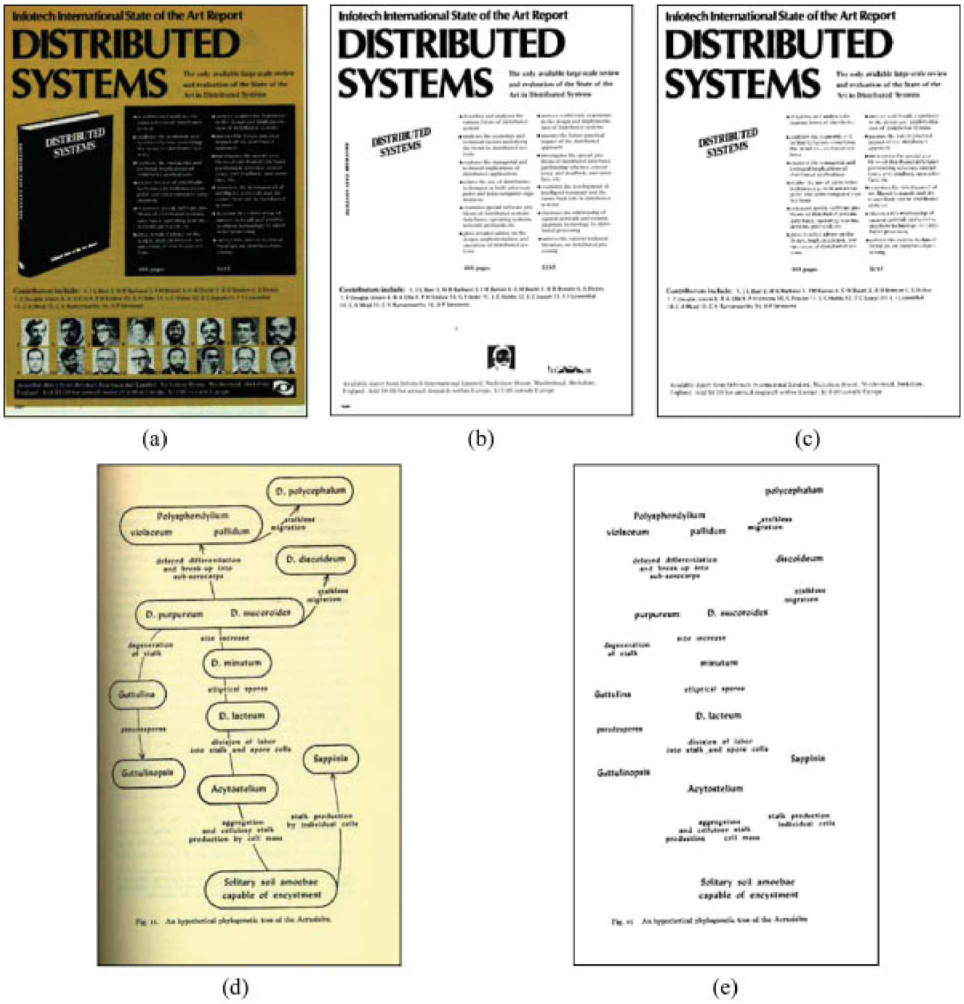


Figure 16. The results obtained by applying to the document of Figure 15(a) the binarization techniques of (a) Otsu, (b) Sauvola, (c) IIFA and (d) ALLT.


Infotech International State of the Art Report
DISTRIBUTED SYSTEMS
Infotech International State of the Art Report
DISTRIBUTED SYSTEMS
Infotech International State of the Art Report
DISTRIBUTED SYSTEMS
D. polycephalum
Polysphondylium
violaceum
pallidum
stalkless migration
delayed differentiation and break-up into sub-sorocarps
D. discoideum
stalkless migration
D. purpureum
D. mucoroides
degeneration of stalk
size increase
D. minutum
Guttulina
elliptical spores
D. lacteum
pseudospores
division of labor into stalk and spore cells
Sappinia
Guttulinopsis
Acytostelium
aggregation and cellulose stalk production by cell mass
stalk production by individual cells
Solitary soil amoebae capable of encystment
Fig. 11. An hypothetical phylogenetic tree of the Acrasiales.
polycephalum
Polysphondylium
violaceum
pallidum
stalkless migration
delayed differentiation and break-up into sub-sorocarps
discoideum
stalkless migration
purpureum
D. mucoroides
degeneration of stalk
size increase
minutum
Guttulina
elliptical spores
D. lacteum
pseudospores
division of labor into stalk and spore cells
Sappinia
Guttulinopsis
Acytostelium
aggregation and cellulose stalk production by cell mass
stalk production by individual cells
Solitary soil amoebae capable of encystment
Fig. 11. An hypothetical phylogenetic tree of the Acrasiales.

(a) (b) (c)

(d) (e)

Figure 17. (a), (b), (c) the original advertisement image, the binary image obtained using $T_p = 0.75$ and the binary image obtained using $T_p = 0.9$, (d), (e) the original linedraw image and the binary image obtained using $T_p = 0.75$. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

D. Experiment 4. In this last experiment, the proposed technique is applied to document images coming from the Mediateam Oulu Document Database (Sauvola and Kauniskangas, 1999). Two document images (one advertisement and one linedraw) with the results obtained are shown in Figure 17. The advertisement image is binarized using two different values for the threshold T_p. In the first case, the proposed technique is applied with $T_p = 0.75$, which is the value used in all previous experiments. Figure 17b shows the binary image obtained. As it can be seen there are two image regions that are wrongly recognized as characters. This problem can be solved by increasing the value of the threshold T_p to 0.9. Figure 17c presents the binary advertisement image obtained, using the specific threshold value. The linedraw image (Fig. 17d) is binarized successfully using $T_p = 0.75$, as it can be seen in Figure 17e.

Using the Otsu binarization technique in the final stage, the processing time in a Pentium 4 CPU 2.4 GHz is calculated equal to 7.39 s for the advertisement image (1897 × 2587 pixels) and 4.18 s for the linedraw image (1452 × 2395 pixels).

VIII. CONCLUSIONS

In this article, a new technique for the binarization of identified text blocks in color document images is presented. The proposed technique is based on a color reduction procedure suitable for document images and on a text localization technique for color documents. Initially, the document image colors are reduced to a small number and then a text localization procedure is applied to each color plane. In the identified text blocks, the ACR technique is locally applied. As a final stage, the text blocks are converted to have black characters in a white background. The technique is suitable for complex cover pages and any type of color documents that contain text and graphics highly mixed with the background. It can also be applied efficiently in noisy color document images. The results obtained by the application of the proposed technique to a large number of data set (500 color text images collected from Internet and other 500 scanned from color book covers and journals), most of which are color cover pages, are very encouraging.

REFERENCES

E. Badekas and N. Papamarkos, A system for document binarization, Third International Symposium on Image and Signal Processing and Analysis (ISPA), Rome, 2003.

J. Bernsen, Dynamic thresholding of grey-level images, Proceedings of Eighth International Conference on Pattern Recognition, Paris, 1986, pp. 1251–1255.

W.Y Chen and S.Y. Chen, Adaptive page segmentation for color technical journals' cover images, Image Vis Comput 16 (1998), 855–877.

D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis, IEEE Trans Pattern Anal Mach Intell 24 (2002), 603–619.

L. Fletcher and R. Kasturi, A robust algorithm for text string separation from mixed text/graphics images, IEEE Trans Pattern Anal Mach Intell 10 (1988), 910–918.

H. Hase, T. Shinokawa, M. Yoneda, and C.Y. Suen, Character string extraction from color documents, Pattern Recogn 34 (2001), 1349–1365.

A.K. Jain and S. Bhattacharjee, Text segmentation using Gabor Filters for automatic document processing, Mach Vis Appl 5 (1992), 169–184.

A.K. Jain and Y. Zhong, Page segmentation using texture analysis, Pattern Recogn 29 (1996), 743–770.

K. Jung and J. Han, Hybrid approach to efficient text extraction in complex color images, Pattern Recogn Lett 25 (2004), 679–699.

K. Jung, K.I. Kim, and A.K. Jain, Text information extraction in images and video: A survey, Pattern Recogn 37 (2004), 977–997.

W. Niblack, An introduction to digital image processing, Englewood Cliffs, NJ, Prentice Hall, 1986, pp. 115–116.

N. Nikolaou, E. Badekas, N. Papamarkos, and C. Strouthopoulos, Text localization in color documents, International Conference on Computer Vision Theory and Applications, Setúbal, Portugal, 2006, pp. 181–188.

N. Nikolaou and N. Papamarkos, Color segmentation of complex document images, International Conference on Computer Vision Theory and Applications, Setúbal, Portugal, 2006, pp. 220–227.

L. O'Gorman, The document spectrum for page layout analysis, IEEE Trans Pattern Anal Mach Intell 15 (1993), 1162–1173.

N. Otsu, A thresholding selection method from gray-level histogram, IEEE Trans Syst Man Cybern C Appl Rev 8 (1979), 62–66.

N. Papamarkos, Color reduction using local features and a SOFM neural network, Int J Imag Syst Technol 10 (1999), 404–409.

N. Papamarkos, A. Atsalakis, and C. Strouthopoulos, Adaptive color reduction, IEEE Trans Syst Man Cybern B Cybern 32 (2002), 44–56.

J. Sauvola and H. Kauniskangas, Media team document database II, a CD-ROM collection of document images, University of Oulu, Finland.

J. Sauvola and M. Pietikainen, Adaptive document image binarization, Pattern Recogn 33 (2000), 225–236.

A. Simon, J.C. Pret, and A.P. Johnson, A fast algorithm for bottom-up layout analysis, IEEE Trans Pattern Anal Mach Intell 19 (1997), 273–277.

K. Sobottka, H. Kronenberg, T. Perroud, H. Bunke. Text extraction from colored book and journal covers, Int J Doc Anal Recogn 2 (2000), 163–176.

C. Strouthopoulos and N. Papamarkos, Multithresholding of mixed type documents, Eng Appl Artif Intell 13 (2000), 323–343.

C. Strouthopoulos, N. Papamarkos, and A. Atsalakis, Text extraction in complex color documents, Pattern Recogn 35 (2002), 1743–1758.

C. Thillou and B. Gosselin, Color binarization for complex camera-based images, Proceedings of the Electronic Imaging Conference of the International Society for Optical Imaging (5667), Color imaging X: Processing, hardcopy, and applications,2005, pp. 301–308.

O.D. Trier and T. Taxt, Improvement of 'Integrated Function Algorithm' for binarization of document images, Pattern Recogn Lett 16 (1995), 277–283.

B. Wang, X.-F. Li, F. Liu, and F.-Q. Hu, 2005. Color text image binarization based on binary texture analysis, Pattern Recogn Lett 26 (2005), 1650–1657.

Y. Yang and H. Yan, An adaptive logical method for binarization of degraded document images, Pattern Recogn 33 (2000), 787–807.

Y. Zhong, K. Karu, and A.K. Jain, Locating text in complex color images, Pattern Recogn 28 (1995), 1523–1535.